A Comparative Study of Breast Cancer Diagnosis based on Neural Network Ensemble via Improved Training Algorithms

A Comparative Study of Breast Cancer Diagnosis based on Neural Network Ensemble via Improved Training Algorithms

Hamed Azami[1], *Student Member, IEEE*, and Javier Escudero[1], *Member, IEEE*

***Abstract*— Breast cancer is one of the most common types of cancer in women all over the world. Early diagnosis of this kind of cancer can significantly increase the chances of long-term survival. Since diagnosis of breast cancer is a complex problem, neural network (NN) approaches have been used as a promising solution. Considering the low speed of the back-propagation (BP) algorithm to train a feed-forward NN, we consider a number of improved NN trainings for the Wisconsin breast cancer dataset: BP with momentum, BP with adaptive learning rate, BP with adaptive learning rate and momentum, Polak–Ribikre conjugate gradient algorithm (CGA), Fletcher-Reeves CGA, Powell–Beale CGA, scaled CGA, resilient BP (RBP), one-step secant and quasi-Newton methods. An NN ensemble, which is a learning paradigm to combine a number of NN outputs, is used to improve the accuracy of the classification task. Results demonstrate that NN ensemble-based classification methods have better performance than NN-based algorithms. The highest overall average accuracy is 97.68% obtained by NN ensemble trained by RBP for 50%-50% training-test evaluation method.**

I. Introduction

In women, one of the most common diagnosed cancers and the prevalent reasons of cancer-related deaths worldwide is breast cancer. Study on diagnosis and treatment of this kind of cancer has become an imperative and significant issue for the scientific community [1, 2]. If breast cancer is correctly detected in an early stage, localized tumors can be treated well before the cancer spreads. Therefore, early detection is the first important step to reduce breast cancer mortality [2, 3].

After a breast tumor is detected, it needs to be identified as benign or malignant. It can be considered as a 2-class classification problem in machine learning. Nowadays, machine learning and computing approaches are used to aid the physician in diagnosis of a wide range of diseases [3, 4]. A large number of approaches have been proposed for breast cancer diagnosis using Wisconsin breast cancer database (WBCD) in literature [3-7].

In machine learning, neural network (NN), inspired by biological NN, is a powerful tool to solve non-linear problems [8]. A well-known type of NN is feed-forward NN, in which each artificial neuron has weights assigned to it, with its inputs coming from neurons in the previous layer, and its output is passed to the next layer after processing. Multilayer perceptron (MLP) is a broadly used class of feed-forward NN [9]. There are many methods to train an MLP and each of which has its own advantages and disadvantages [8].

Although the backpropagation (BP) is the most well-known and popular algorithm to train an MLP, it has two main limitations: (1) the BP training is too slow, especially in real-time applications, and (2) the BP easily falls in a local minimum. To overcome these limitations, BP with momentum, BP with adaptive learning rate, BP with adaptive learning rate and momentum, four kinds of conjugate gradient algorithms (CGAs), including Polak–Ribikre CGA, Fletcher-Reeves CGA, Powell–Beale CGA, and scaled CGA, resilient BP (RBP), one-step secant (OSS) and quasi-Newton are employed here to train an MLP. It should be mentioned that several of these algorithms are used for the dataset such as [10-12]. However, no study covers all mentioned approaches.

Ensemble methods have recently been used in classification techniques with great success [13, 14]. The combinational methods frequently have higher robustness, accuracy, and resistance than single classification methods. The motivation for this technique is based on the intuitive idea that, by combining the outputs of a number of individual predictors, the performance of a single one will more likely enhance [13-15].

The rest of the paper is organized as follows. The dataset is briefly described in Section II. The concept of NN ensemble and various training algorithms are presented in Section III. In Section IV, results and discussion are presented. Finally, conclusions are explained in Section V.

II. Materials

In this piece of research, we use a publicly available breast cancer dataset obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [16] and taken from UCI machine learning repository (http://archive.ics.uci.edu/ml/). It includes 699 instances taken from fine needle aspirates of human breast tissue. Since 16 instances of the dataset have missing information, 683 instances are used in our experiment, including 444 and 239 instances respectively belong to benign (not harmful) and malignant (may be harmful) classes. Each instance has nine attributes as shown in Table I [16].

III. Methods

A. Neural Network ensemble

NN ensemble is a relevant topic in machine learning and data mining [13, 17]. It is generally accepted that a

[1] Hamed Azami * and Javier Escudero are with the Institute for Digital Communications, School of Engineering, The University of Edinburgh, Edinburgh, King's Buildings, EH9 3JL, United Kingdom. (Phone: +44 131 650 5599, emails: hamed.azami@ed.ac.uk, javier.escudero@ed.ac.uk). * *Corresponding author.*

combination of numerous different prediction approaches can improve predictions. For combination of a number of results, we may need 1) to employ different clustering/classification algorithms to produce partitions for combination, 2) to change initialization or other parameters of clustering/classification methods, 3) to use different features via feature extraction for subsequent clustering/classification, and 4) to partition various subsets of the original data [18]. For this purpose, we employ the ensemble of NN by the use of changing the initial weights of an MLP.

In our paper, for each training algorithm, initial weights are randomly assigned in the first step. Then, an NN is trained using the initial weights. It should be mentioned that 50% of the dataset, randomly selected, is used for training and the rest is used for testing each algorithm. After training the NN, the test part of the dataset is used to calculate the output of 50% of the test part. Next, this procedure is repeated 10 times. Inasmuch as the initial weights are changed in every repetition, the results may differ. The final stage is finding a method to compound their results and make final decision. There are some approaches to combine some results [19]. Combination method of classification results is generally dependent on their output type. For soft/fuzzy outputs, majority vote, simple average, and weighted average are three well-known approaches to ensemble [14, 19]. In simple averaging method, the average of outputs is calculated and then the class that has the highest average value is chosen as the final decision. Weighted averaging procedure is like simple average, except that a weight for each classifier is used for calculating that average. Majority vote is one of the most common combining techniques which is based on soft/fuzzy outputs. In this method, the combiner votes to class *j* if a little more than half of base classifiers vote to class *j* [13]. Here, the majority vote is used to combine the results.

B. Neural Network Training Algorithms

Albeit BP is a widespread method to train an MLP, the convergence of this method is too slow, because it is mainly a steepest descent algorithm [8, 20, 21]. To tackle this limitation, we employ BP with momentum, BP with adaptive learning rate, BP with adaptive learning rate and momentum, Polak–Ribikre CGA, Fletcher-Reeves CGA, Powell–Beale CGA, scaled CGA, RBP, OSS, and quasi-Newton method.

TABLE I. THE LIST OF NINE ATTRIBUTES OF THE BREAST CANCER DATASET

Attributes	Domain	Mean	Standard deviation
Clump Thickness	1-10	4.42	2.82
Uniformity of Cell Size	1-10	3.13	3.05
Uniformity of Cell Shape	1-10	3.20	2.97
Marginal Adhesion	1-10	2.80	2.86
Single Epithelial Cell Size	1-10	3.21	2.21
Bare Nuclei	1-10	3.46	3.64
Bland Chromatin	1-10	3.43	2.44
Normal Nucleoli	1-10	2.87	3.05
Mitoses	1-10	1.59	1.71

The learning rate, the scale of the increments of the weight at every updating step, can significantly affect the performance of the training algorithm for a feed-forward NN. A large learning rate value may cause instability while a very small learning rate can slow down the training procedure. To overcome this limitation, adaptive learning algorithm was employed [8, 22]. In this method, when the change in the sum of squared errors has the same algebraic sign for several consequent epochs, the learning rate parameter goes up while the change in the sum of squared errors has the different algebraic signs for several consequent epochs, the learning rate parameter goes down [21, 22].

In the basic BP, the weights in the steepest descent direction (negative of the gradient, the direction in which the performance function is decreasing most quickly), are adjusted. It is worth noting that albeit the error function reduces most quickly along the negative slope of the gradient, it does not unavoidably create the fastest convergence. In the CGAs, a search is done along conjugate directions, which generally makes a faster convergence in comparison with that of the steepest descent direction [8, 22]. There are four kinds of CGAs, namely, Polak–Ribikre, Fletcher-Reeves, Powell–Beale, and scaled CGA. Although each of first three CGAs has its own benefits and drawbacks, their concept are relatively similar. Scaled CGA, proposed by Moller, uses a step-size scaling mechanism that avoids a time-consuming line search per learning iteration [23]. These algorithms are described in detail in [8, 21].

Newton's method is an alternative to the CGAs for fast optimization. This algorithm frequently converges faster than CGAs, yet it is time-consuming and complex to compute the Hessian matrix for a feed-forward NN. There are a number of algorithms based on Newton's method, which do not require to calculate the second derivatives. These methods, named quasi-Newton, update an estimate of the Hessian matrix at each iteration of the algorithm and then, the update is computed as a function of the gradient [8, 20, 22].

One shortcoming of the quasi-Newton algorithm is that updated parameters are required to be storage for a matrix of size $N*N$ and calculations are of order $O(N^2)$. Albeit the available storage is less of a problem now than it was in the past, the computational problem still exists when N is too large. An alternative is to use a secant approximation with O(N) computing. In the OSS method , a new search direction is considered from vectors calculated from gradients [8, 20, 22].

The RBP is an appealing algorithm for supervised learning in feed-forward NNs. It is an improved static NN and known to provide faster local adaptation of weights and biases without sacrificing accuracy. The RBP, which is a first order optimization algorithm, is a high speed algorithm to converge in the defined space [21, 24].

IV. Results and Discussion

In the first step, the dataset is randomly divided into two equal training and test data sets. In the second step, we employ an MLP with three layers and, for all approaches, we train the NN with training dataset and then use the test dataset for evaluating these algorithms. The momentum and initial learning rate are respectively set to 0.85 and 0.05. Due to the uncertain behavior of the NNs, we run all methods 40 times, and the average of the results is presented.

The number of neurons plays a key role in the performance of an NN. In case the number of neurons of hidden layer is too low, the NN may not able to model complex data and the resulting may be unreasonable, whereas choosing a large number increases the training time as well as may reduce the performance of the NN. Considering this fact, we chose the number of neurons of the hidden layer as many as 20 by some trials. Moreover, we changed the number of iterations from 10 to 1000. When we increase the number of iterations to more than 1000, the accuracy of each of the methods does not remarkably change. However, the training time goes up considerably. The activation function is another effective parameter of an MLP. Linear, tangent sigmoid and logarithm sigmoid are three widely used functions in an NN. We selected the logarithm sigmoid function for the NN by using trial and error.

Table II illustrates the classification accuracies using different algorithms for training the MLP for diagnosis of breast cancer. As can be seen in Table II, for BP with momentum, BP with adaptive learning rate, and BP with adaptive learning rate and momentum, the accuracies first increase significantly until iteration number of 500 and then, decrease slightly. It is worth noting that, as expected theoretically [8, 22], BP with adaptive learning rate and momentum have better performance than BP with adaptive learning rate and BP with adaptive learning rate is better than BP with constant momentum.

Among CGAs, scaled CGA has the highest accuracies and the other three ones have fairly similar performance in all iterations, although their accuracies initially increase moderately and then decrease slightly. A very clear difference between BP-based ones and CGAs is that in iteration 10, the accuracies obtained by the first approaches are considerably lower than the latter algorithms. With only 10 iteration, the accuracy of NN-based method using scaled CGA is 0.9149. Thus, scaled CGA may be employed in real-time applications.

The classification results obtained by quasi-Newton training algorithm fluctuate from 0.8964 to 0.9175. The accuracy of OSS-based NN first increases temperately and after iteration number of 500 it decreases slightly. Among these NN classifiers, for diagnosis of breast cancer, RBP is the best algorithm to train an MLP almost in each iteration. The accuracy of NN-based approach trained by OSS first increases until iteration 500 then, decreases slightly. It is worth to note that after RBP and scaled CGA, OSS has the highest classification accuracies in low iterations.

Table III demonstrates the accuracies obtained by different training algorithms for NN ensemble. As expected, the combination of results improves the accuracies reported in Table II. As mentioned in [13], when the classification accuracy of a method is too low, its ensemble may not change this accuracy. This fact can be seen in BP with momentum and BP with adaptive learning with iteration 10. To compare Tables II and III, it is clear that the majority of accuracies in Table III are higher than those corresponding accuracies in Table II. It shows the importance of ensemble technique. Like Table II, Table III shows that RBP is superior to the other training algorithms for an NN ensemble. It should be added the best training algorithms for NN ensemble, like NN, are RBP, scaled CGA, and OSS in low iterations.

When comparing Tables II and III, it is worth to mention that the training time for NN ensemble is about 10 times more than that for NN, because in NN ensemble approach, we repeat each method 10 times and then combine them. Thus, regarding running time for training the NN, every method in Table II can be relatively equaled to its corresponding one with tenfold more iterations.

In Table IV, the best algorithms among the abovementioned ones, named RBP-based NN and RBP-based NN ensemble in Tables II and III, respectively, are compared with four existing well-known methods. As can be seen, the accuracy of RBP-based NN ensemble is superior to the other algorithms. It is worth noting that because a 10-fold cross validation is used in many papers, we cannot report their results although the dataset is similar. Since in 10-fold cross validation, in each step, 90% of instances for training and only 10% of them for test are used, their results are usually higher than those of using 50%–50% training-test evaluation approach. This fact may be motivated because having higher number of samples in the training set usually leads to better performance [25].

Table II. Comparison of averages of accuracies (40 iterations) of different algorithms training an MLP for diagnosis of breast cancer (50%–50% training-test). The number of iterations varies from 10 to 1000.

Methods	10	50	100	200	500	1000
BP with momentum	0.5304	0.5330	0.6322	0.6863	0.7699	0.7431
BP with adaptive learning rate	0.5355	0.7426	0.8463	0.9632	0.9604	0.9589
BP with adaptive learning rate and momentum	0.5633	0.5982	0.8907	0.9598	0. 9713	0.9618
Fletcher-Reeves CGA	0.8955	0.8921	0.9224	0.9262	0.9258	0.9172
Powell-Beale CGA	0.8700	0.9065	0.9104	0.9175	0.9315	0.9260
Polak-Ribiére CGA	0.8678	0.9063	0.9176	0.9237	0.9281	0.9257
Scaled CGA	0.9149	0.9557	0.9493	0.9450	0.9476	0.9553
Quasi-Newton	0.8964	0.9045	0.9175	0.9044	0.9123	0.9118
OSS	0.9015	0.9075	0.9089	0.9220	0.9574	0.9447
RBP	0.9683	0.9685	0.9616	0.9591	0.9584	0.9583

TABLE III. COMPARISON OF AVERAGES OF ACCURACIES (40 ITERATIONS) OF DIFFERENT ALGORITHMS TRAINING AN NN ENSEMBLE FOR DIAGNOSIS OF BREAST CANCER (50%–50% TRAINING-TEST). THE NUMBER OF ITERATIONS VARIES FROM 10 TO 1000.

Methods	10	50	100	200	500	1000
BP with momentum	0.5269	0.6140	0.7368	0.8158	0.9339	0.9290
BP with adaptive learning rate	0.4468	0.9368	0.9708	0.9713	0.9662	0.9515
BP with adaptive learning rate and momentum	0.6211	0.8667	0.9708	0.9719	0.9567	0.9556
Fletcher-Reeves CGA	0.9737	0.9620	0.9602	0.9585	0.9573	0.9573
Powell-Beale CGA	0.9731	0.9591	0.9550	0.9585	0.9561	0.9579
Polak-Ribiére CGA	0.9743	0.9596	0.9579	0.9544	0.9591	0.9573
Scaled CGA	0.9737	0.9626	0.9556	0.9561	0.9567	0.9567
Quasi-Newton	0.9719	0.9702	0.9620	0.9641	0.9597	0.9669
OSS	0.9737	0.9731	0.9643	0.9591	0.9597	0.9561
RBP	0.9737	0.9760	0.9678	0.9620	0.9643	0.9667

TABLE IV. COMPARISON OF AVERAGES OF ACCURACIES OF THE BEST EMPLOYED ALGORITHMS TRAINING AN NN ENSEMBLE-BASED AND SEVERAL WELL-KNOWN EXISTING METHODS FOR DIAGNOSIS OF BREAST CANCER (50%–50% TRAINING-TEST).

Methods proposed in	RBP-NN	RBP-NN Ensemble	[6]	[26]	RBF [27]	PNN [27]
Correct classification rate	0.9685	0.9768	0.9655	0.9589	0.9618	0.97

V. CONCLUSIONS

In this piece of research, several improved NN trainings have been used for classification of the Wisconsin breast cancer dataset. We have also used the ensemble NN concept, via changing initialization of each NN-based approach, to enhance their performance. Results show that NN ensemble-based classification approaches have had better performance than NN-based ones and the best training algorithm has been RBP to classify the dataset. We intend to improve other classification approaches, such as support vector machine, using different ensemble techniques. We will also combine some results obtained by classifiers with different improved training algorithms.

REFERENCES

[1] K. Mungrue, J. Ramdath, S. Ali, W.-A. Cuffie, N. Dodough, M. Gangar*, et al.*, "Challenges to the Control of Breast Cancer in A Small Developing Country," *Breast cancer: basic and clinical research,* vol. 8, p. 7, 2014.

[2] H. Gewefel and B. Salhia, "Breast cancer in adolescent and young adult women," *Clinical breast cancer,* vol. 14, pp. 390-395, 2014.

[3] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Systems with Applications,* vol. 42, pp. 4611-4620, 2015.

[4] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network," *Expert Systems with Applications,* vol. 38, pp. 9573-9579, 2011.

[5] F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," *Pattern Analysis and Applications,* pp. 1-10, 2014.

[6] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications,* vol. 38, pp. 9014-9022, 7// 2011.

[7] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research,* vol. 43, pp. 570-577, 1995.

[8] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural network design*: Pws Pub. Boston, 1996.

[9] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines* vol. 3: Pearson Education Upper Saddle River, 2009.

[10] F. M. Al-Naima and A. H. Al-Timemy, *Resilient back propagation algorithm for breast biopsy classification based on artificial neural networks*: INTECH Open Access Publisher, 2010.

[11] R. Setiono and L. C. K. Hui, "Use of a quasi-Newton method in a feedforward neural network construction algorithm," *Neural Networks, IEEE Transactions on,* vol. 6, pp. 273-277, 1995.

[12] F. Paulin and A. Santhakumaran, "Classification of breast cancer by comparing back propagation training algorithms," 2011.

[13] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems,* pp. 231-238, 1995.

[14] M. Mohammadi, H. Alizadeh, and B. Minaei-Bidgoli, "Neural network ensembles using clustering ensemble and genetic algorithm," in *Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on*, 2008, pp. 761-766.

[15] H. Azami, J. Escudero, A. Darzi, and S. Sanei, "Extracellular spike detection from multiple electrode array using novel intelligent filter and ensemble fuzzy decision making," *Journal of Neuroscience Methods,* vol. 239, pp. 129-138, 1/15/ 2015.

[16] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the national academy of sciences,* vol. 87, pp. 9193-9196, 1990.

[17] M. Pulido, P. Melin, and O. Castillo, "Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange," *Information Sciences,* vol. 280, pp. 188-204, 2014.

[18] A. Azadeh, M. Saberi, M. Anvari, and M. Mohamadi, "An integrated artificial neural network-genetic algorithm clustering ensemble for performance assessment of decision making units," *Journal of intelligent manufacturing,* vol. 22, pp. 229-245, 2011.

[19] C. Dietrich, G. Palm, and F. Schwenker, "Decision templates for the classification of bioacoustic time series," *Information Fusion,* vol. 4, pp. 101-109, 2003.

[20] H. Azami and S. Sanei, "GPS GDOP classification via improved neural network trainings and principal component analysis," *International Journal of Electronics,* vol. 101, pp. 1300-1313, 2014.

[21] H. Azami, M.-R. Mosavi, and S. Sanei, "Classification of GPS satellites using improved back propagation training algorithms," *Wireless personal communications,* vol. 71, pp. 789-803, 2013.

[22] H. Demuth and M. Beale, "Neural network toolbox for use with MATLAB," 1993.

[23] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural networks,* vol. 6, pp. 525-533, 1993.

[24] P. Mastorocostas, "Resilient back propagation learning algorithm for recurrent fuzzy neural networks," *Electronics Letters,* vol. 40, pp. 57-58, 2004.

[25] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.

[26] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing,* vol. 17, pp. 694-701, 2007.

[27] T. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," *Journal of electrical & electronics engineering,* vol. 4, pp. 1149-1153, 2004.